International Journal of scientific research and management (IJSRM)
\||Volume||3||Issue||3||Pages|| 2457-2460||2015||
Website: www.ijsrm.in ISSN (e): 2321-3418

Data Deduplication: An approach for avoidance of duplicate data

G.Karthika[1], M.P.Ruby[2], S.Vijalakshmi[3], M.Purushothaman[4] , S.Anbarasan[5]

School of Engineering and Technology, Surya Group of Institutios[1,2,3,4,5],
Villupuram, India,
kkgopaalan@gmail.com[1],m.p.rubyramesh@gmail.com[2],vijiit90@gmail.com[3],purushoth.89@gmail.com[4],anbu.179206@gmai.com[5]

Abstract: ***Cloud computing is an emerging technology; it provides many services regarding software, infrastructure, storage, etc. Storage as a service is very eminent service and makes users to use their storage space resourcefully. In cloud computing, to reduce the utilization of storage area and bandwidth data deduplication technique is used. This technique eliminates the redundancy data. To provide semantic security for the data, a new scheme is projected. By sorting out the data as encrypted and un-encrypted data can attain semantic security than existing system. To improve the rate, reliability and accessibility of the data backup scheduling is proposed.***

Keywords: deduplication, semantic security, bandwidth

1. Introduction

Cloud computing is the emerging technology which provides many services such as software-as-a-service, infrastructure-as-a-service, platform-as-a-service. Cloud computing is the delivery of computing services over the internet. This cloud computing also provides storage-as-a-service. Cloud services are popular because users can access their e-mail, social networking site or photo service from anywhere in the world, at any time, at minimal or no charge. Some cloud providers may, use the personal information of users for advertising purpose or to learn more about the users for the others reasons. Now-a-days cloud services providers provides highly available storage as well as parallel computing resources in minimum costs. We are focusing on the cloud storage service, which offers a good advantage to the users to store their data or files in remote storage. The cloud user can upload and download their data at anytime and anywhere by connecting to the cloud storage through any devices. Here the data stored in the cloud storage can be same in various devices. The tremendous growth of the data volumes burdens the cloud storage. To minimize the consumption of cloud storage data deduplication technique is used. Data deduplication is a technique which avoids the redundancy data and stores only one copy of them in the cloud storage. Because of this technique the cloud storage volume can be used efficiently.

In cloud storage, users concentrate about the personal privacy of the data which they are storing in it. Personal privacy is referred as data confidentiality. Users encrypt their files and store them in the cloud storage. But there is always a conflict between data deduplication and data confidentiality. For example, consider two users, user A and user B. Both the users need to store their data in the cloud storage where their data is identical. Here, user B needs to encrypt the file because he/she needs some privacy whereas user A needs not to encrypt the data, without encrypting he/she is storing the data in the cloud storage. The data deduplication is not applicable when the data is encrypted.

To avoid this conflict between data deduplication and data confidentiality convergent encryption is used, that make data deduplication feasible. The basic form of convergent encryption is taking the original file and calculating a hash from it. Then using this hash as the key, encrypt the rest of the file. This function is 100% deterministic, so any clients encrypting the same data will generate the same hash key and encrypted data. This convergent encryption never provides semantic security of the data stored by the cloud user in the cloud data. To overcome this problem, it maintains two types of cloud storage as encrypted and un-encrypted data.

There are lots of applications in the cloud backups. The approach of backup is to attain the data reliability and data availability even at any natural disaster or any sort of system failure. Now-a-days, data volume scheduled by the cloud backup scheduling server is increasing along with the time.

The only expectation of the cloud user will be that cloud backup scheduling server should fulfill the speed and reliability of data backup. The cloud backup scheduling server should check about the state details of every storage node namely data access speed, task load, free storage room, etc.

In order to increase the cloud backup speed, we introduce two parameter of scheduling decision, data access time and data reliability. To avoid the drawback of dumping the data in a single storage node and also to avoid the problem of overflow of data in the cloud storage node, this cloud backup scheduling is introduced.

2. Related Work

Fatema Rashid et al [1], 2012, proposed that, Cloud Storage Provider does not need to be trusted when handling the user's data. It has also been designed to ensure privacy of the data by providing only encrypted copy of data to Cloud Service Provider. The major limitation of this work is, however they are using Convergent Encryption Algorithm to encrypt the chunks. The encryption key of a convergent encryption is the hashed value of a plaintext. Therefore it also does not satisfy confidentiality. Also the collision of hash has not been considered here.

Twin Clouds Architecture. Recently, Bugiel et al. provided an architecture consisting of twin clouds for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud. Zhang et al. also presented the hybrid cloud techniques to support privacy-aware data-intensive computing. In our work, we consider to address the authorized deduplication problem over data in public cloud. The security model of our systems is similar to those related work, where
the private cloud is assume to be honest but curious.

Convergent Encryption. Convergent encryption ensures data privacy in deduplication. Bellare et al. formalized this primitive as message-locked encryption, and explored its application in space-efficient secure outsourced storage. Xu et al. also addressed the problem and showed a secure convergent encryption for efficient encryption, without considering issues of the key-management and block-level deduplication. There are also several implementations of convergent implementations of different convergent encryption variants for secure deduplication. It is known that some commercial cloud storage .providers, such as Bitcasa, also deploy convergent encryption.

3. METHODOLOGY

As convergent encryption has drawback of lack of providing semantic security of the data. To avoid this drawback in the cloud storage, it maintains two types of cloud storage as encrypted and un-encrypted data. One is an un-encrypted data (i.e.) data will be in plaintext format. Another cloud storage is an encrypted data where all data are encrypted (i.e.) in ciphertext format. In our proposed system, a message (i.e.) the data or files which is send by the cloud user consists of three blocks they are hash check block, enabling block, ciphertext block.

In the hash check block, it computes the hash value of file and it also checks the repetition of the stored files. Here the hash value is encrypted with the public key of the cloud server and the resultant hash value is stored in this block. In the enabling block, it generates an AES key k (session key) and it encrypts k with the public key of the user, finally the resulted key is stored in this block. In ciphertext block, encrypt the file and its hash value with the AES key k, the resultant hash value is stored in this block. These are the structure of the three blocks. This backup scheduling has two components such as cloud state table and cloud resource table.

3.1 Mechanism for plaintext deduplication:

- For uploading data, User A sends a data block to the server S. Now, S checks the hash value of this data block or file by comparing it with has list which is maintained by the server.
- There are two cases, in case 1: if no duplication occurs, S stores the received file and add the hash value in the hast list. In case 2: if duplication occurs, it deletes the received file sent by the cloud user and jus store the link which denotes to the first copy of the duplicate data block.
- For downloading data, User A request to download the data block. In case 1: if the data block is not a link then it returns the data. In case 2: if the data is a link, then it returns the linked data block from the storage node.

3.2 Mechanism for ciphertext deduplication:

- For uploading the data, User B sends a triple of blocks to server S. Now S, obtains the hash value and it also the checks it by comparing it with the hash list.
- In case 1: if no duplication occurs, store the received triple of blocks. In case 2: If duplication occurs, the enabling block content is replaced with the null symbol and the cipher text block content is replaced with the link which denotes to the first copy of the duplicate data block.
- For downloading the data, User B requests to download the data. In case 1: if the enabling block content is not a null symbol then return the triple of blocks. In case 2: if the content of the enabling block is a null symbol. After renewing the ciphertext block, S randomly selects an AES key and with the help of this key it encrypts the linked data block. Then renewing the enabling block, S takes User b's public key of the cloud server to encrypt the AES key.

Decryption of Ciphertext Deduplication: In ciphertext deduplication decryption of key is carried out. User B decrypts the enabling block with the help of his/her private key and then takes the AES key to decrypt the ciphertext block.

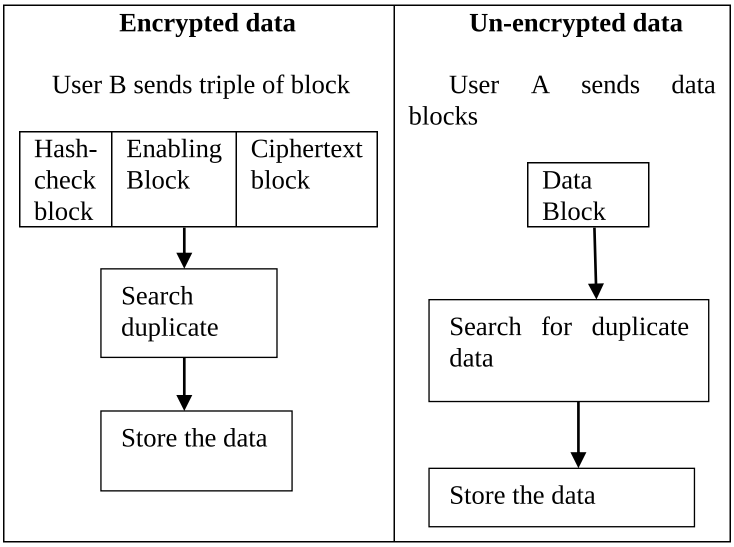

Figure 1: Splitting up of data

3.3 Tables of cloud backup scheduling

Cloud State Table (CST) describes the state information of each storage nodes in cloud. Cloud Resource Table (CRT) is designed to record how data are stored. There are two parameters for scheduling decision viz., data access time and data reliability. In this work the reliability of data is guaranteed by two copies of each file. The storage nodes in cloud are distributed to different places and the probability of two storage nodes are damaged at the same time is little. Therefore, if one copy of the data was destroyed by some kind of disaster, the other redundant copy of the data can still help in recovery [6]. The cloud backup scheduling architecture is given below,. The CST contains information of cloud storage nodes such as IP address, total disk space, available disk space, disk access speed and number of storage request in the queue. The network latency was omitted in this approach. The CRT contains information that is used to identify which data stored in which storage node. The information in CRT are data identifier and data centers name. The data identifier denotes the hash value of the file. If the data is stored in a particular storage node then it will be set as one, otherwise the value will be set as zero.

When the user sent request to the backup scheduling server the server will update the CST waiting time by calculating the file size divided by disk access speed. After updating CST the server will find out the decision parameter, which is used to backup the data. The decision parameter will be decided by finding the load balancing and time taken for input and output operation of each storage node. The load balancing factor is calculated as a ratio of free disk to total disk space of storage node. The I/O operation time is calculated by adding waiting time and disk access speed of particular storage node. After the calculation of decision parameter, the two values which are minimum will be taken. According to these parameters the data will be backup into the particular storage node and then while completion of backup the CST value will be updated by reducing the free disk space.

This work could be further improved by following techniques.

- The redundancy could be avoided by data de-duplication technique.
- The security could be provided to user's data.
- The storage space could be efficiently used by compressing the file before storing into server.

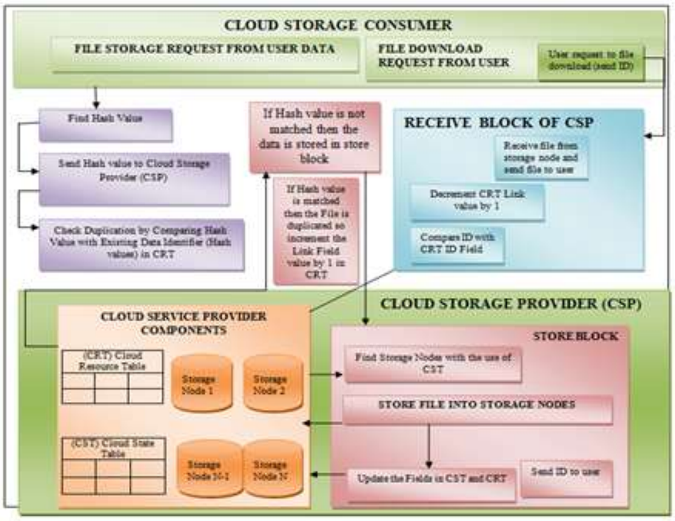
CLOUD STORAGE CONSUMER
FILE STORAGE REQUEST FROM USER DATA
FILE DOWNLOAD REQUEST FROM USER
User request to file download (send ID)
Find Hash Value
Send Hash value to Cloud Storage Provider (CSP)
Check Duplication by Comparing Hash Value with Existing Data Identifier (Hash values) in CRT
If Hash value is not matched then the data is stored in store block
If Hash value is matched then the File is duplicated so increment the Link Field value by 1 in CRT
RECEIVE BLOCK OF CSP
Receive file from storage node and send file to user
Decrement CRT Link value by 1
Compare ID with CRT ID Field
CLOUD STORAGE PROVIDER (CSP)
CLOUD SERVICE PROVIDER COMPONENTS
(CRT) Cloud Resource Table
Storage Node 1
Storage Node 2
(CST) Cloud State Table
Storage Node N-1
Storage Node N
STORE BLOCK
Find Storage Nodes with the use of CST
STORE FILE INTO STORAGE NODES
Update the Fields in CST and CRT
Send ID to user

Figure 2: Architecture for proposed system

3.4 Cloud Storage Provider (CSP)

The Cloud Storage Provider provides the storage space to user to store their files. This is a private cloud storage server. The users can store their files into storage server and they can retrieve the files when they need it. The functionality of Cloud Storage Provider is discussed below.

Components of Cloud Storage Provider

The Cloud Storage Provider has three components with it. They are Cloud State Table (CST), Cloud Resource Table (CRT), and Storage Nodes. The CST has Storage node ID, IP Address, Total disk space, Free disk space, Disk access speed, and Wait time. CRT has Data Identifier, Link, and Storage node field.

Store Block

The store block is used to store files into storage nodes. The store block is used to store the file which is not duplicated. Store block finds the storage nodes n1 and n2 using the decision parameter.

Receive Block

When the users need to download the data which was stored by them into cloud storage server, the user should send the ID which was provided by the Cloud Storage Provider. The receive block compares this ID with the CRT Data Identifier field values. If the value matches then it will retrieve the file from the corresponding storage node and sends the file to user. After that, receive block decrements the corresponding link field value by one. If we search file in every storage node, it will take more time to search. But, this is an efficient way to receive the block, because the storage node which is having the file has been founded by comparing ID with CRT ID field values. So, the time is reduced very much.

3.5 Cloud Storage Consumer

Cloud storage consumer consumes the storage space from CSP. This is a pay as you use technology. The cloud storage consumer can store their files into CSP and retrieve the stored files from CSP. There are two requests from cloud storage consumer to cloud storage provider. They are explained briefly below.

File Backup Request

The user sends file storage request to CSP when they need to backup their files. The file may be encrypted or unencrypted format. The store block in CSP is used to store the users file into storage nodes.

File Download Request

The user sends file receive request to CSP when they need to download their files. The user sends ID which has been sent by CSP. This ID was provided by CSP when the users backup their files. The receive block in CSP is used to receive the users file into storage nodes.

3.5 Backup Request

The user sends file storage request to CSP when they need to backup their files. The file may be encrypted or unencrypted format. If the file is unencrypted then, the hash value will be calculated for that file and then this hash value will be sent to the CSP. The CSP will compare this hash value with CRT data Identifier field values. If the hash value matches then, the link value is incremented by one. Otherwise, the file will be transferred to CSP and the file will be stored using store block.

If the file is encrypted then, the hash value which is stored in check block will be sent to CSP. The CSP will compare this hash value with CRT data Identifier field values. If the hash value is matched then, it just increments the link field value by one. If the hash value matched then, it just increments the link field value by one and enabling block and cipher block will be made as NULL and stored into CSP using store block. Otherwise, the triple block (check block has hash value of the file, enabling block has key which is encrypted using user's public key, and cipher block has file which is encrypted using key) will be transferred to CSP.

According to the time t update the wait time of the CST using the equation given in (4.4). If any storage node is damaged then the corresponding waiting time of that storage node will be set to infinity and it indicates that, the storage node could not be reached. Select the storage nodes whose available disk space is greater than the size of the file and the waiting time should be less than infinity. These selected storage node list will be in set T. The two storage nodes which will be used to backup the file determined. The storage node n1 and n2 are selected through the decision parameter value from the set of nodes in T. The selected storage node decision parameter value should be minimum comparing than other storage nodes decision value in set T. After finding the storage nodes, compress the file using gzip. This compression will be used to reduce the storage space consumed by the file. Then backup the compressed file into storage node n1 and n2. The ID and backup address of data will be inserted into CRT. The ID will be sent to the user.

3.7 Download Request

The user sends file receive request to CSP when they need to download their files. The user sends ID which has been sent by CSP when the users backup their files. The receive block receives the file from storage node by comparing the hash values in CRT with the ID sent by user. If the file is single block then the file will sent to user. If the file is triple block then it will be sent to user. Otherwise, the file is triple block and then enabling block and cipher block is marked as NULL then, the CSP will generate the key and encrypts the file using this key and then stores into cipher block. The key will be encrypted using private key of server and stores into enabling block. The check block has hash value of the file. These triple blocks will be sent to user. The user will decrypt the key using server's public key and then the file will be decrypted using the key.

4. Results & Discussions

Calculate that how much time it takes when a cloud user downloads a file from the encrypted area. The renewing process of the enabling block and ciphertext block takes place only when the null symbol is there in respective blocks. The server must execute an AES encryption while renewing the cipher block. An AES encryption operation needs 1003 milliseconds with a 100-megabyte file and a 16-byte AES key on a CPU-GPU platform. So these computations are quite efficient.

4.1 PERFORMANCE METRICS

The two performance metrics are considered for proposed IBSD system namely, bandwidth and storage utilization ratio.

4.4.1 Bandwidth

The IBSD system determines the duplicate of the data at client side. The hash value of the client data are compared with the hash values which is stored in CRT. If the hash value matches then the data will not be sent to the storage provider. Thus, it reduces the consumption of bandwidth.

4.4.2 Storage Utilization Ratio (SUR)

The new metric called Storage Utilization Ratio is proposed here. The Storage Utilization Ratio is defined as ratio between the storage space utilized without de-duplication and storage space utilized with de-duplication.

5. Conclusion

The cloud computing is used to provide everything as a service on demand basis. The Storage-as-a-Service is provided by Cloud Storage Provider which is used to backup the user data. The concept of cloud computing is presented and also techniques available in cloud computing is also reviewed. The cloud computing provides many advantage to the consumer. It offers everything as a service to consumer an on-demand basis.

An overview of various existing backup scheduling is presented. The existing system concentrates on backup speed and reliability, but it does not provide de-duplication technique. The reliability in existing system is provided by making copy of the same data into two storage nodes. The security is also not provided to the user data.

The de-duplication technique is very important to reduce the consumption of storage space. The de-duplication technique is used to avoid the data redundancy. There are many technologies available for data de-duplication. The techniques of de-duplication such as SIS, WFC, FLC, and CBC are explained in literature review briefly. The existing de-duplication approaches avoid the redundancy of files using the concept of convergent encryption scheme. The convergent encryption algorithm uses hash value of data as the encryption key. So, the service provider can easily extract the data. In this situation the confidentiality will not be provided.

The issue in existing de-duplication techniques is solved by the proposed system which handles two types of data. One is encrypted data and another one is unencrypted data. The de-duplication will be provided in two various strategies for these two types of data. The proposed work handles the de-duplication without compromising the availability and scheduling speed. Also the proposed system will provide the semantic security to user data.

References

[1] Fatema Rashid, Ali Miri, Isaac Woungang, " A Secure Data De-duplication Framework for Cloud Environments", Tenth Annual International Conference on Privacy, Security and Trust, 2012.

[2] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In *ASIACCS*, pages 195–206, 2013.

[3] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In *Technical Report*, 2013.

[4] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui. A secure cloud backup system with assured deletion and version control. In *3rd International Workshop on Security in Cloud Computing*, 2011.

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In *EUROCRYPT*, pages 296–312, 2013.

[6] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In *Proc. of APSYS*, Apr 2013.